# OPTIMIZING STRATIFICATION WITH DATA-DRIVEN APPROACH: A CASE STUDY

Ilaria Bombelli, Giorgia Sacco

**Abstract.** The stratified sampling design is helpful in official statistics to guarantee the accuracy and efficiency of survey estimates. The choice of suitable stratification variables and the subsequent determination of the strata are two of the most important steps in this procedure. When the stratification variables are continuous, to create the strata, these variables must be converted into categorical variables that identify classes.

The quality of the stratification can be greatly impacted by the division of these continuous variables, specifically the choice of intervals or class boundaries. Class intervals are sometimes predetermined based on established procedures or past knowledge. In some cases, however, the researcher must determine the best partitions into classes based on the population's characteristics and the survey's goals. Choosing such partitions can be a hard task.

The R package `SamplingStrata` provides useful assistance in situations when the researcher has the flexibility to define the stratum. This tool is based on a genetic algorithm and provides a data-driven approach for determining the best stratification boundaries, thereby maximizing sampling design efficiency while satisfying precision requirements.

In this study, we present an application of the aforementioned package on a household survey dataset. We illustrate the advantages of utilizing the package to guide the stratification process, especially when working with continuous auxiliary variables. The findings demonstrate how, compared to conventional techniques based on arbitrary or fixed class definitions, data-driven stratification can result in more efficient sample allocations.

## 1. Introduction

Stratified sampling is one of the most widely used sampling strategies, particularly in official statistics. The population is divided into distinct groups, known as strata. It is desirable for these to be internally homogeneous and externally heterogeneous to optimize the efficiency of the design and to have accurate estimates. To create such strata, available information from population-level auxiliary variables is exploited to improve the precision of survey estimates while keeping data collection costs under control. However, the performance of this approach is highly dependent on how the strata are defined. This is particularly crucial when the auxiliary variables involved are continuous in nature, and determining how to discretize them effectively becomes a methodological challenge.

In institutional surveys, strata based on continuous variables are often constructed using a priori defined class intervals, whose thresholds are often defined through expert heuristic rules or by administrative cut-offs. While rather simple to implement operationally, a fixed class definition for continuous variables may fail to capture the underlying distributional characteristics of the data. In addition, strata may result to be suboptimal and potentially inefficient since they could lead to obtain a final allocation of a rather high sample size in order to achieve the precision level goals set (Sarndal *et al*., 2003).

As an alternative to a priori defined class intervals, a data-driven approach to find the cutoff values to determine the classes can be adopted. In the literature, several proposals have been discussed. Khan et al. (2008) introduced a dynamic programming method to determine the optimal strata boundaries that minimize the overall variance under Neyman allocation. Subsequently, Khan and Sharma (2015) used nonlinear mathematical programming to jointly optimize both the boundaries and the sample sizes between strata. In the field of integer linear programming, Wright (2014) introduced exact allocation algorithms in the context of stratified sampling that incorporate integral constraints and layer-specific bounds. We propose to address this challenge by using the `SamplingStrata` package in R (Barcaroli, 2014). This package uses a genetic algorithm to search for optimal cut-off points in the continuous auxiliary variables to minimize the sample size and achieve the set precision levels. The algorithm integrates this with the Bethel-Chromy allocation procedure, ensuring efficient sample distribution across strata under linear modelling assumptions.

To evaluate the effectiveness of this approach, we compared two stratification strategies: one that defines standard strata by discretizing the continuous variable a priori and the other that uses the boundaries obtained through a data-driven approach implemented in the `SamplingStrata` procedure. The effectiveness of the proposed method is evaluated both in terms of the coefficients of variation obtained for the same sample size and in terms of the sample size required to achieve a predefined precision target.

The paper is structured as follows. Section 2 presents the methodological framework and algorithms used. Section 3 illustrates the application of the case study on household survey data in both the standard procedure and the data-driven approach. Section 4 summarizes the main results of this work.

## 2. Methodology

To determine both the optimal population stratification and the corresponding allocation, subject to predefined precision constraints, `SamplingStrata` package in R (Barcaroli, 2014; Barcaroli *et al.*, 2020), provides a flexible tool in this framework. Indeed, its flexibility lies in the fact that it allows for management of the stratification determination when both categorical and continuous variables are used.

We recall that when choosing auxiliary variables for stratification, the researcher should focus on variables highly correlated with the target variable to improve the performance, both in terms of reducing the coefficient of variation in the domains of interest and in terms of reducing the sample size.

For this reason, when a continuous variable is a target variable, we can also use it as an auxiliary variable for stratification. This scenario represents the ideal use case since there is a perfect correlation between the auxiliary variable and the target variable (Lavallée, 1988).

In real-world surveys, both categorical and continuous variables are considered as stratification variables. To stratify using both categorical and continuous variables, when a-priori classes of the continuous variable are not provided, a data-driven approach can help. We need to define the domains by cross-classifying the categorical variables, and then for each domain, we need to discretize the continuous variable. Then, the final stratification is the combination of the domains and the cuts of the continuous variable found for each domain. In this way, we are able to construct a stratification framework that reflects the complexity of a design often encountered in real-world surveys (Särndal *et al.*, 2003).

To implement this strategy, we use the `SamplingStrata` R package. The first step in the procedure is to construct a sampling frame using the `buildFrameDF()` function. Using this function, each unit in the population is associated with a unique identifier, auxiliary variables, and domain membership.  Then, within each domain, stratification is driven by the continuous variable that needs to be discretized.

To determine the number of classes into which to divide the continuous variable within each domain, the `KmeansSolution2()` function is applied. This function uses the k-means algorithm to provide insight into the possible optimal number of final classes and to provide an initial solution for the optimization phase. Next, the `optimStrata()` function is applied with the method set to "*continuous*" (specific for continuous-type variables). This function applies a genetic algorithm to find the optimal combination of cut points in the continuous variable. Each candidate solution corresponds to a specific stratification that is evaluated through a fitness function that estimates the total sample size required to satisfy the coefficient of variation (CV) constraints and chooses the combination of cuts that minimizes the required sample size. In addition, the function also calculates the sample allocation between strata using Bethel's algorithm (Bethel, 1989).

## 3. Application

To test the data-driven approach to determine the stratification, we use the population dataset provided by the R2BEAT package[1], filtered on individuals aged over 14, which contains demographic and economic information on 2050016 population units. It's worth mentioning that this dataset is a synthetic dataset that mimics a real population dataset. For each individual, identified by an ID (id_ind), we have information about his/her residence (region/province/municipality), the ID of the family he/she belongs to, his/her gender, his/her age in aggregated form (class of age), his/her level of education, his/her income, and three flags indicating whether or not he/she is employed, unemployed, inactive (1: yes, 0: no). In Table 1 a complete description of the available variables is provided.

**Table 1 –** *Variables' descriptions.*

| Variable | Description | Type |
|---|---|---|
| id_ind | Identifier of individual | Categorical (2050016 categories) |
| reg | Region of residence | Categorical (3 categories) |
| prov | Province of residence | Categorical (6 categories) |
| mun | Municipality of residence | Categorical (513 categories) |
| id_fam | Identifier of individual's family | Categorical (961457 categories) |
| gender | Gender of individual | Categorical (2 categories) |
| Cleta | age class: 14-24; 24-34; 34-44; 44-54; 54-64; 64-74; 74+ | Categorical (7 categories) |
| Titstu | Level of Education | Categorical (5 categories) |
| Income | Income of the individuals | Numeric |
| Employed | Flag indicating individual is employed | Categorical (2 categories) |
| Unemployed | Flag indicating individual is unemployed | Categorical (2 categories) |
| Inactive | Flag indicating individual is inactive | Categorical (2 categories) |

We suppose we want to implement a sampling survey aiming at measuring the population income, and the estimates are given at the following domains of

[1] https://github.com/barcaroli/R2BEAT/tree/master/data, dataset pop.RData

estimates: DOM1: national level; DOM2: province; DOM3: age class; DOM4: province crossed with age class. To design a sample, we opt for a stratified sampling design and we choose as stratification variables age class (Cleta), province (prov), and income, as we believe they can be correlated with the target variable.

Following a standard approach, the continuous variable is manually discretized as a new categorical variable (clred, which stands for class of income), with 6 categories: 0-4999, 5000-9999 10000-19999, 20000-29999, 30000-99999, 100000+. The final stratification is given by crossing the categories of province, age class, and income class, for a total of 252 strata.

To determine the final allocation in stratified sampling designs, several methods are available. Assume that $n$ units are to be allocated among H strata, with $\boldsymbol{n_h}$ units allocated among each stratum $\boldsymbol{n_h}, h \in \{1, \dots, H\}$. Clearly, $\sum_{\{h=1\}}^{H} \boldsymbol{n_h} = \boldsymbol{n}$. The allocation of units inside strata can be uniform, proportional, and optimal. The uniform allocation allows allocating the same constant number of units in each stratum; proportional allocation allows allocating more units in strata which have a higher dimension; finally, optimal allocation allows allocating more units in strata which have a higher dimension and are characterized by a higher variability of the target variable (s). The optimal allocation with one target variable and one domain has been introduced by Tschuprow (1923) and Neyman (1934); its extension to the multivariate framework is thanks to Bethel (1989); finally, Falorsi *et al.* (1998) developed the optimal allocation in a multivariate, multi-domain framework. Ballin and Barcaroli (2013) developed a methodology for the joint determination of the best stratification and best allocation in a multivariate and multi-domain framework.

Given the aforementioned strata, the optimal allocation of units in a multidomain and multivariate framework has been determined with the function `beat.1st` of the `R2BEAT` R package (Barcaroli *et al.*, 2023). The final sample consists of n = 10000 units.

Conversely, following a data-driven approach, we use the `SamplingStrata` R package to make the stratification. We decided to make the package discretize the variable income, for each combination of the categories of province (prov) and age class (cleta) (i.e., a total of 42 combinations). In detail, we use equal precision constraints in all domains (42 combinations), set to 0.01, and the algorithm is configured with 50 iterations, with parallel computation enabled to improve efficiency.

Among the outputs, one describes for each combination of province's and age class' categories, the boundaries of the cuts of the income variable, the stratum label, the number of population units falling in the stratum, and the number of sampling units allocated in the stratum. Table 2 shows the aforementioned output of the first six strata for the domain prov:1_cleta:(14,24].

**Table 2 –** *Summary of 6 strata of Domain "prov:1_cleta:(14,24]" Stratification.*

| Domain | Stratum | Population (Nstr) | Allocation (SOLUZ) | Lower Income | Upper Income |
|---|---|---|---|---|---|
| prov:1_cleta:(14,24] | 1 | 18859 | 61 | 0.000 | 4572.43 |
| prov:1_cleta:(14,24] | 2 | 7584 | 13 | 4573.130 | 7018.21 |
| prov:1_cleta:(14,24] | 3 | 12070 | 42 | 7019.220 | 12087.49 |
| prov:1_cleta:(14,24] | 4 | 3438 | 5 | 12089.390 | 14050.20 |
| prov:1_cleta:(14,24] | 5 | 6664 | 27 | 14052.290 | 19909.03 |
| prov:1_cleta:(14,24] | 6 | 4853 | 33 | 19911.350 | 30002.83 |

The two approaches to stratification are discussed in terms of results evaluation, focusing on both the estimates' precision and the final sample size.

## 4. Results evaluation

The first analysis carried out to compare results obtained by the standard approach and the data-driven approach to stratification is devoted to the Coefficient of Variations (CV), which are a measure of estimates' errors. They are used to measure the precision of the estimates and are obtained by computing the ratio between the standard deviation of the estimates to the mean; obviously, the lower the coefficient of variation, the more precise the estimate; viceversa, the higher the coefficient of variation, the less precise the estimate. Therefore, we compute for each estimate domain the expected coefficient of variation of the income. This is done both for the stratification and allocation obtained with a standard approach and for the stratification and allocation obtained with a data-driven approach.

To make results comparable, we adjust the resulting allocation from the data-driven approach to stratification to make the final sample size equal to the one obtained from the resulting allocation corresponding to the standard approach to stratification (n=10000): the adjustment is made proportionally within each stratum and preserves the optimality of the solution (Bethel, 1989).

Figure 1 shows a radar plot reporting the expected coefficient of variation of every category of all the domains of interest for the allocations obtained with both a standard approach to stratification and a data-driven (proposed) approach to stratification. As it can be easily seen, the proposed approach leads to lower CVs in every domain than the ones obtained with a standard approach. Table 3 summarizes the results of the radar plot, by reporting for every target domain, the maximum level of expected coefficient of variations obtained with both the standard and data-driven (proposed) approach. From the table, it is possible to appreciate that the proposed approach performed better than the standard one.

**Figure 1 –** *CV for each domain category. Comparison between standard and proposed stratification with fixed final sample size.*

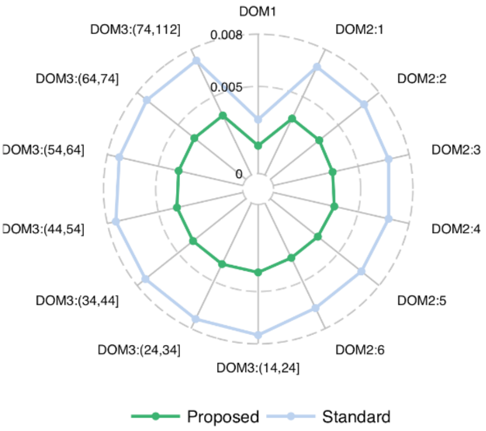
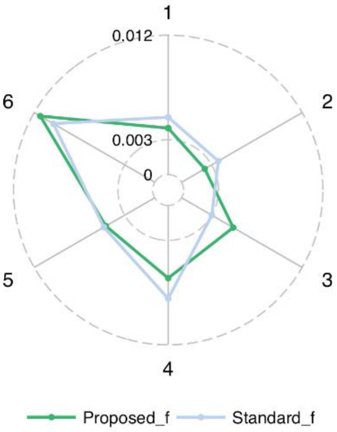


**Table 3 –** *Max CV (%) for each domain. Comparison between standard and proposed stratification with fixed final sample size.*

| Domain | Max CV (%) standard allocation | Max CV (%) proposed allocation |
|---|---:|---:|
| DOM1 | 0.31 | 0.16 |
| DOM2 | 0.69 | 0.36 |
| DOM3 | 0.75 | 0.39 |
| DOM4 | 1.62 | 0.65 |

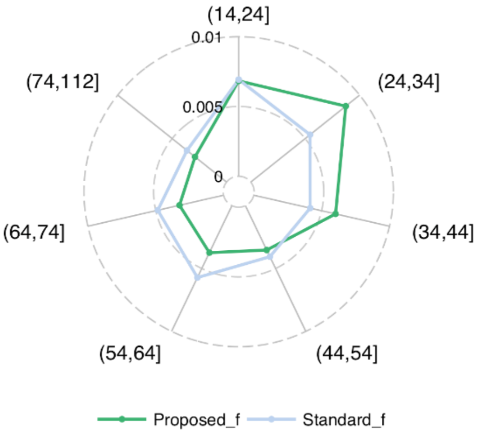
Given that the final sample size is the same in both approaches (n=10000), we also investigate the allocation separately in each domain. The radar plot represented in Figure 2 shows the sampling fraction in each domain category of province (DOM2). Only in province 3 and 6, the sampling fractions of the proposed approach are higher than the standard one; however, they are balanced by the lower sampling fractions in the other categories of the domain, and the larger amount of sampling units in these categories of the domain leads to a lower level of CV.

**Figure 2 –** *Sampling fraction for each domain category of DOM2 (province). Comparison between standard and proposed stratification with fixed final sample size.*



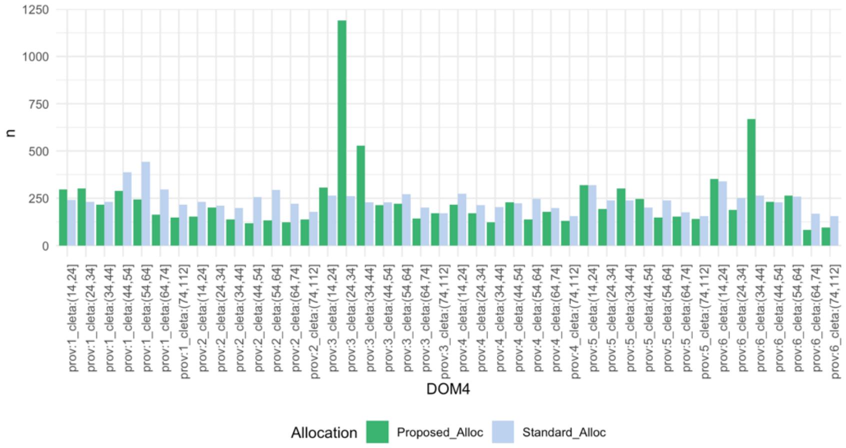
The radar plot represented in Figure 3 shows the sampling fraction in each domain category of age class (DOM3). Only in age classes (24, 34] and (34, 44] the sampling fractions of the proposed approach are higher than the standard one: however, they are counterbalanced by the other categories' smaller sampling fractions, and the higher number of sampling units in these categories results in a lower CV.

**Figure 3 –** *Sampling fraction for each domain category of DOM3 (age class). Comparison between standard and proposed stratification with fixed final sample size.*



Finally, the bar plot displayed in Figure 4 shows the comparison between allocations in each domain category of domain province crossed with age class (DOM4). It is possible to notice that the proposed approach allows to obtain an allocation lower than the one obtained with a standard approach in most of the domain categories. However, we notice that the proposed allocation is strongly higher than the standard allocation in domains prov:3_cleta:(24,34], prov:3_cleta:(34,44], prov:6_cleta:(34,44]. Results are coherent with the sampling fractions showed in the previous radar plots (Figure 2 and Figure 3).

**Figure 4 –** *Final allocation for each domain category of DOM4 (province crossed with age class). Comparison between standard and proposed stratification with fixed final sample size.*

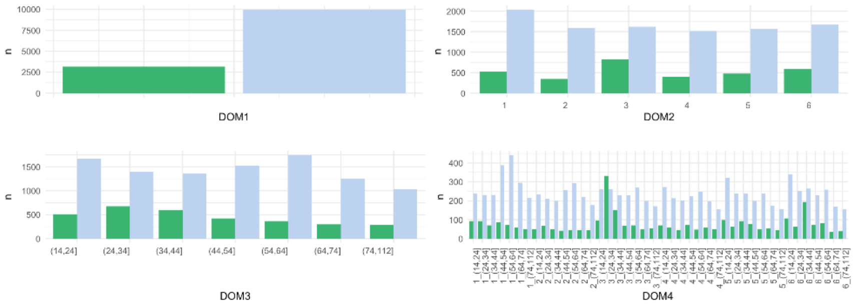


Therefore, we can conclude that in terms of estimates' precision, with the same final sample size, the proposed approach allows for obtaining more precise estimates, with coefficient of variations lower than the one obtained with the standard approach in all the domains of interest.

We implemented another analysis on the comparison of the results. So far, we have shown that, by keeping the final sample size fixed, the estimates are better in terms of CV. We now decide to test whether, keeping the CV fixed equal to that of the standard approach, the proposal still preserves its better performance in terms of final sample size. Fixing the CV, we run the data-driven proposed approach to determine the stratification and the allocation. The final allocation with the proposed approach consists of 3162 sampling units. Therefore, we can conclude that the proposal preserves its good performance also in terms of sample size. Indeed, a lower sample size allows for a reduction in the overall survey cost.

The overall sample size obtained with the proposed approach (3162) is significantly lower than the one obtained with the standard approach (10000), and we can also appreciate it separately by domains. As it is highlighted in Figure 5, the final proposed allocation is always significantly lower than the standard one in all the categories of the domains of interest. The only exception is province 3, age class (24, 34], where the proposed allocation is slightly above.

**Figure 5 –** *Final allocation for each domain category of DOM1 (national level), DOM2 (province), DOM3 (age class), DOM4 (age class crossed with province). Comparison between standard and proposed stratification with fixed CV.*



## 5. Conclusions

When dealing with a stratified sampling design, two main aspects must be considered: first, the choice of the stratification variables; then, the way numerical stratification variables are divided into classes.

In this paper, we showed how it is possible to adopt a data-driven approach to help the researcher in the definition of the classes using the `SampingStrata` R package.

It is important to underline that the package is primarily designed for use with real-world data. The synthetic dataset we used in the application closely resembles real-world data, with around 2 million records; results suggest that the package can handle large datasets efficiently. To further speed up computation, the parallel=TRUE option can be used for parallel processing, reducing execution time.

In terms of package functions, most parameters need to be defined by the researcher arbitrarily, depending on the specific problem (e.g., population frame, error constraints, domains, target variables, auxiliary variables). However, some parameters require further tuning, such as the number of iterations and the number of strata. The package documentation provides default values and suggests using a k-means algorithm to initialize these parameters.

The proposal turns out to be helpful once we compare its performance with that obtained with standard classes definition, i.e. classes defined a priori. The performance is measured both in terms of expected coefficient of variations, by keeping the final sample size fixed, and in terms of final sample size, by keeping the coefficient of variations fixed. In both cases, the performance of the proposed approach is significantly better than the standard approach: the coefficient of variations are halved and the final sample size is reduced by about 70%, from 10000 to 3162 sampling units.

## References

BARCAROLI G., BALLIN M., ODENDAAL H., PAGLIUCA D., WILLIGHAGEN E., ZARDETTO D. 2020. *Sampling Strata: optimal stratification of sampling frames for multipurpose sampling surveys*, R package.

BARCAROLI G. 2014. SamplingStrata: An R package for the optimization of stratified sampling. *Journal of Statistical Software*, Vol. 61, pp. 1–24.

BARCAROLI G., FASULO A., GUANDALINI A., TERRIBILI M. D. 2023. Two Stage Sampling Design and Sample Selection with the R Package R2BEAT. *The R Journal*, Vol. 15, No. 3, pp. 191–213.

BETHEL J. 1989. Sample allocation in multivariate surveys. *Survey methodology* Vol. 15, No. 1, pp. 47–57.

COCHRAN W. G. 1977. *Sampling techniques*. John Wiley & Sons.

FALORSI P. D., BALLIN M., DE VITIIS C., SCEPI G. 1998. Principi e metodi del software generalizzato per la definizione del disegno di campionamento nelle indagini sulle imprese condotte dall'ISTAT. *Statistica Applicata*, Vol. 10, No. 2, pp. 235–257.

KHAN M.G.M. NIRAJ N. NURAIN A. 2008. Determining the optimum strata boundary points using dynamic programming. *Survey methodology*. Vol. 34, pp. 205-214.

KHAN M. G., SHARMA S. 2015. Determining optimum strata boundaries and optimum allocation in stratified sampling. *Aligarh Journal of Statistics*, Vol. 35, pp. 23-40.

LAVALLÉE P. 1988. Two-way optimal stratification using dynamic programming.*Proc. Sect. Surv. Res. Methods*, pp. 646-651.

NEYMAN J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society,* Vol. 97, No. 4, pp. 558–625.

SÄRNDAL C. E., SWENSSON B., WRETMAN, J. 2003. *Model assisted survey sampling*. Springer Science & Business Media.

TSCHUPROW A. A. 1923. On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron,* Vol. 2, pp. 646–683.

WRIGHT T. 2014. A simple method of exact optimal sample allocation under stratification with any mixed constraint patterns. *Statistics,* Vol. 7.

Ilaria BOMBELLI, Istat, ilaria.bombelli@istat.it
Giorgia SACCO, Istat, giorgia.sacco@istat.it